NOTES 21: CLT-BASED INFERENCE FOR DIFFERENCES

Stat 120 | Fall 2025

Prof Amanda Luby

0.1 CLT Recap

CLT for 1 Proportion

$$\hat{p} \sim N(\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_)$$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

But, we don't know p, so plug in __________ (CI's), __________ (HT's), __________ (conservative)

CLT for 1 Mean

$$\bar{X} \sim N(\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_)$$

where $SE_{\bar{X}} =$ __________. But, we don't know σ, so we plug in s. Then,

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Example We want to build a 99% confidence interval for μ, where $\bar{X} = 290$, $n = 120$ and $s = 87.6$. The correct form for the CI is:

0.2 CLT for a difference in proportions

Idea:

Population parameter:

Sample statistic:

CLT:

$$\hat{p}_1 - \hat{p}_2 \sim N(\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_)$$

How big is "big enough?"

- $n_1 p_1 >$ __________

- $n_2 p_2 >$ ___________

What is the SE?

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

but, we don't know p_1 or p_2. Three options:

1. "Plug-in" $\hat{p}_1$ and $\hat{p}_2$
2. Use $\hat{p}_{pooled} =$ _________________
3. "Conservative" SE: assume $p =$ _________________

CLT-based confidence interval for a difference in proportions:

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE_{plug-in}$$

CLT-based test for a difference in proportions:
$H_0 : p_1 - p_2 = 0$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{pooled}}$$

0.3 CLT for a difference in means

Idea:

Population parameter:

Sample statistic:

CLT:

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, SE_{\bar{x}_1 - \bar{x}_2})$$

How big is "big enough?"

Case 1:

- $n_1 >$ __________________
- $n_2 >$ __________________

Case 2:

- $n_1 \leq$ _____ or $n_2 \leq$ _____
- AND

What is the SE?

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

but, we don't know σ_1 or σ_2, so we plug in s_1 and s_2 instead:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

the sampling distribution is no longer a normal distribution, but becomes a ____________________

CLT-based confidence interval for a difference in means:

$$\bar{x}_1 - \bar{x}_2 \pm t^*_{df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

CLT-based test for a difference in means: $H_0 : \mu_1 - \mu_2 = 0$

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $df =$ _____________________

0.4 CLT for Matched Pairs

Idea:

Population parameter:

Sample statistic:

CLT:

How big is "big enough"?

What is the SE?

CLT-based confidence interval for a matched pairs difference:

CLT-based test for a matched pairs difference: $H_0 : \mu_d = 0$

1 Group Problems: Difference in Proportions

APM Research Lab ran a survey of 800 likely Minnesota voters between Sept 16-18 of this year. They reported 48.4% were in support of the Harris/Walz ticket, and 43.3% were in support of the Trump/Vance ticket.

Here's an excerpt from their methodology report:

> *The margin for error, according to standards customarily used by statisticians, is no more than ± 3.5 percentage points. The margin of error is approximately 7 percentage points for the difference between two data points…. The margin of error is higher for any subgroup, such as gender or age grouping.*

(a) What is the *standard error* for the difference in proportions?

(b) Explain why we should *not* use the CLT formula for the difference in proportions between Harris supporters and Trump supporters in this poll

(c) Instead, we'll look at the difference in support for Harris between two different levels of education. (We can assume that the poll included 436 in the "no college degree" group and 360 in the "college degree" group). Show two ways that the standard error could have been computed.

(d) The two ways of computing the standard error should produce similar results. Give two reasons why.

(e) The following R code makes a confidence interval and performs a hypothesis test for the difference in these two proportions.

i. Where did `x = c(192, 194)` come from?

ii. What is α for the hypothesis test?

iii. Provide an in-context interpretation of the confidence interval (it should be clear in your answer which group had more support for Harris)

```r
prop.test(x = c(192, 194),
          n = c(436, 360),
          correct = FALSE, #tells R not to apply a "continuity correction"
          conf.level = .90)
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(192, 194) out of c(436, 360)
X-squared = 7.7, df = 1, p-value = 0.006
alternative hypothesis: two.sided
90 percent confidence interval:
 -0.15680 -0.04024
sample estimates:
prop 1 prop 2
0.4404 0.5389
```